面向高校学生微博的跨粒度情感分析 *

刘　丽，岳亚伟

(山西农业大学 软件学院，山西 太谷 030801)

摘　要：传统的微博情感分析往往忽略不带感情色彩的情感词对微博情感的影响，并缺乏对复杂句式的分析。为此，提出结合条件随机场（conditional random filed，CRF）和复杂句式的跨粒度情感分析方法。该方法在 CRF 模型的基础上，融合复杂句式特征和语义依存特征，对学生微博进行细粒度情感分析，识别出微博文本中的情感要素，在此基础上，通过基于复杂句式的粗粒度情感分析方法分析微博文本的情感倾向，实现对学生总体情感倾向的跨粒度分析。实验结果显示，跨粒度情感分析方法的提出，使得情感要素识别的综合准确率达 88%左右，微博情感分析的综合准确率达 87%左右。比起传统的情感分析方法，准确率更高，分类效果更好。

关键词：高校学生微博；条件随机场；复杂句式；跨粒度；情感分析

中图分类号：TP391.1　　doi: 10.3969/j.issn.1001-3695.2017.12.0815

Cross-grained sentiment analysis oriented to college student micro-blog

Liu Li, Yue Yawei

(*School of Software Shanxi Agricultural University*, *Taigu Shanxi* 030801, *China*)

Abstract: Traditional sentiment analysis of micro-blog often ignore the influence of sentiment words that have no sentimental color on micro-blog sentiment, and lack of analysis for complex sentence. To solve the problem, this paper proposed a method of cross-grained sentiment analysis based on conditional random filed and complex sentence, which fused complex sentence and semantic dependency features on the basis of CRF. It can identify sentiment elements by analyzing micro-blog sentiment in fine-grained. The method of coarse-grained sentiment analysis based on complex sentence was used to analyze sentimental tendency of student micro-blog. Finally, the experimental results show that the accuracy on sentiment elements can reach 88%, furthermore, the accuracy of micro-blog sentimental tendency can reach 87%. Compare to traditional method, the method we proposed has higher accuracy and better performance.

Key words: college student micro-blog; conditional random filed(crf); complex sentence ; cross-grained; sentiment analysis

0　引言

随着网络的发展，人们表达情感的方式也逐渐趋于多样化，更多的是通过网络平台，以文本的形式来表达。尤其对于当代的高校大学生，好奇心强，接受新事物的能力强，更容易接受现在较发达的社交平台。通过这些平台他们可以接受到最新的网络信息，包括各种社会新闻、娱乐八卦、前沿科技等，也可以通过这些平台发表意见、表达情感、记录生活点滴。微博就是这样一种社交平台，在这里，学生可以表达自己的情绪和观点，获取学生发布的这些信息能够方便高校管理者更好的了解学生的思想动态，捕捉学生整体的情感倾向。但是如果仅靠人工浏览，无法应对海量的学生微博，因此情感分析技术应运而生。通过对微博文本进行情感分析，可以提取大量有价值信息，分析出学生的喜怒哀乐，对于高校管理者而言，可以获取最新的学生动态，及时与学生沟通，并作出正确的引导。

情感分析按照分析粒度的不同，有粗粒度情感分析和细粒度情感分析，粗粒度情感分析是对篇章和句子进行情感分析，细粒度情感分析是对短语或者词语进行情感分析。对高校学生微博的情感分析属于粗粒度的情感分析，主要包括情感词典法和机器学习法。

情感词典法需要将分词后的文档或句子中的每个词与情感词典中的词进行匹配，并统计匹配成功的正负面情感词的数量，通过数量判断文本的情感倾向。为了能更准确地识别情感词，肖江等人[1]构建了基准情感词典，并在基准情感词典的基础上构建了相关领域情感词典，采用相似度计算的方法确定领域情感词的情感倾向。文献[2]主要是基于中文微博构建情感词典，采用的方法是扩展的点互信息 So-PMI 算法，可以自动获得领域情感词，并加入到基础情感词典中[2]。这类方法从情感词典

收稿日期：2017-12-23；**修回日期：**2018-02-12　　**基金项目：**青年科技创新基金资助项目（2017016）

作者简介：刘丽（1990-），女，山西临汾人，助教，硕士，主要研究方向为文本分析、数据挖掘（sxaulily@163.com）；岳亚伟（1988-），男，讲师，硕士，主要研究方向为机器学习，人工智能．

出发，机械的将分词后的文本与情感词典进行匹配，匹配成功则认为是带有感情色彩的情感词，如："好难过啊，一天的课"。情感词典匹配法就会将"好"和"难过"都标记为情感词，但实际上"好"在文本中是程度副词，修饰情感词"难过"，表达"难过"的程度。显然，情感词典法无法正确匹配不带感情色彩的情感词。因此情感分析的效果往往不理想。

机器学习法，需要标注文本语料，利用机器学习模型训练这些语料，得到文本分类模型。常用支持向量机、朴素贝叶斯、最大熵模型为分类模型。Catal 等人[3]采用了多种分类器进行情感分析，包括朴素贝叶斯算法、支持向量机、Bagging 算法，最后利用投票算法决定分类的最终结果。

Liu 等人[4]将 Co-training 协同训练算法与 SVM 相结合进行推文的情感分析，Co-training 协同训练算法可以实现语料的半自主标注，省时省力，再利用 SVM 算法实现推文的情感分类。

但是机器学习模型，在判断文档和句子的情感倾向时，跟情感词典法一样，极有可能忽略文本中不带感情色彩的情感词，而且缺乏对复杂句式的考虑。因此，需要结合细粒度情感分析，先识别出真正影响文本情感倾向的情感要素，再分析微博文本的情感倾向。

细粒度情感分析可以分析出影响情感表达的各个要素，包括情感对象，正负面情感词以及影响文本情感倾向的结构词。Hu 等人[5]认为评价对象一般是名词或名词短语，评价词一般为形容词，因此采用关联规则法来实现情感要素的抽取。文献[6]也是基于关联规则法，制定了产品属性词与情感词之间对应的词性模板，然后基于该模板提出了一种抽取属性词及其对应情感词的算法，实现了产品属性词和匹配情感词的识别[6]。Xu 等人[7]首先识别出情感句，然后利用句子结构特征和词语搭配关系，抽取出与情感发起者最相关的核心词语，再结合句法特征扩展核心词语，最终情感发起者即为扩展后的最长名词短语。

李阳辉等人[8]提出基于深度学习的细粒度情感分析，不同于文献[5~7]对单一语料的细粒度情感分析，它分析的对象来自不同的语料，包括评价词典、微博、影评、知乎等，分析的粒度从词语级别到篇章级别。

这类方法都属于无监督学习方法，没有充分利用文本语言特征，忽略了词与词之间的语义依存关系。

有监督学习方法通常是利用机器学习模型来进行情感分析，常用的机器学习模型有隐马尔科夫模型和条件随机场模型，Jin 等人[9]提出词汇化的隐马尔科夫（hidden Markov models,HMM）框架进行意见挖掘分析，通过序列标注的方法将与产品相关的各个实体及对实体的相关意见标注出来，从而确定意见对象和意见词，但仅选取了词性和上下文线索为特征，没有充分利用各种语言特征。Liu 等人[10]将条件随机场与最大熵模型相结合进行情感分析，提取词、上下文信息，词的位置等特征，利用条件随机场进行序列标注，提取 unigram、bigram 特征，选用最大熵模型判断整个句子的情感倾向。这类方法都是在机器学习模型的基础上，选取一些语言特征，实现文本的细粒度情感分析，好的特征对情感分析有至关重要的作用。

本文旨在对高校学生微博进行情感分析，需要粗粒度的情感分析方法，但是粗粒度的分析微博句子，缺乏对不带感情色彩的情感词的分析，也忽略了复杂句式对微博情感倾向的影响。因此融合细粒度情感分析，提出跨粒度情感分析方法，细粒度情感分析将复杂句式特征和语义依存特征融入到条件随机场中，可以充分分析不带感情色彩的情感词，识别出微博句子中的真正的正负面情感词，而且还可以识别出影响微博情感倾向的复杂结构词。结合识别出的情感词和复杂结构词，采用基于复杂句式的粗粒度情感分析方法分析微博文本的情感倾向，从而实现粗粒度和细粒度情感分析相互强化的跨粒度情感分析。

1 情感要素识别

1.1 条件随机场模型

条件随机场模型思想主要来源于最大熵模型，2001 年被 Lafferty 等人[11]首次提出，克服了隐马尔科夫模型严格独立假设的要求，可以以序列标注的形式将情感要素提取出来。首先将经过分词的微博文本作为观察序列，如：X={ x_1, x_2, x_3 ......}，x_i 为分词后的词语，将观察序列作为 CRF 模型的输入数据，所有可能的标注状态的条件概率就被计算出来，最后选择条件概率最大的那个标注状态输出，即 Y={ y_1, y_2, y_3 ....}，y_i 为对应的 x_i 的标注状态[12]，具体计算公式如下：

$$P(Y/X)=\frac{1}{Z(X)}\exp(\sum_i\sum_k \lambda_k f_k(y_{i-1},y_i,X,i)) \quad (1)$$

其中：Z(X)为归一化因子，可使求得的概率 P 满足概率要求，计算公式如下：

$$Z(X)=\exp(\sum\sum \lambda_k f_k(y_{i-1},y_i,X,i)) \quad (2)$$

其中：X 为观察序列，即分词后的微博文本，Y 为文本的标注结果序列，$f_k(y_{i-1},y_i,X,t)$ 为任意的布尔型特征函数，λ_k 为特征函数对应的权值，选取 P 最大时的输出序列，即为观察序列的标注状态。

1.2 语料标注

本文选用的语料，是通过爬虫获取的"山西农业大学"学生的相关微博文本，总共 10000 条。情感要素的抽取需要大量标注语料，人工标注的方式可靠却效率低下。故采取文献[13]提出的基于 MapReduce 的协同训练(Tri-training)模型对语料进行半自动标注[13]。具体的标注状态分为四类：正面情感（Positive Sentiment,PS）词，即识别出的词语若为正面情感词，简记为 PS；负面情感（Negative Sentiment,NS）词，即识别出的词语若为负面情感词，简记为 NS；复杂结构（Complicated Structure,CS）词，即识别出的词语若为复杂结构词，简记为 CS；背景（Background word，BW）词，即识别出的词语若为背景词，简记为 BW。详细说明如表 1 所示。

表 1　标注集说明

标注	说明
CS	表明该词为影响微博情感倾向的结构词，如转折词、否定词
PS	表明该词为表达正面情感或正面情绪的词
NS	表明该词为表达负面情感或负面情绪的词
BW	表明该词是 CS，PS，NS 三种词之外的其他词

1.3　扩展情感词典

现在的微博文本语言随意灵活，从而衍生出大量新颖的网络词汇，如“ 无语、醉了、蓝瘦香菇”等词语，一些词语还表现出一定的情感色彩，如“蓝瘦香菇”即为“难受，想哭”的意思，表现出了负面情绪，但现有的情感词典无法匹配日益更新的网络词汇，会影响情感特征的提取，从而影响情感要素抽取的准确性，故提出扩展情感词典的方法，将 Hownet 的中文正负面情感词典作为基础情感词典，利用 NLPIR 汉语分词系统的新词发现功能，抽取出微博文本中的新词，然后再通过 word2vec 模型，分析这些新词的情感倾向，最终确定带有感情色彩的新词，将它们加入到基础情感词典中，具体过程如下：

a）提取新词。NLPIR 系统不仅可以实现自适应分词，还可以从较长的微博文本中，基于信息交叉熵自动提取新词。但该工具本身一次只能分析一条文本，无法处理海量文本。因此利用该工具提供的系统开发文档，找到提取新词的接口程序，稍作修改，即可遍历所有的微博文本来提取新词。

b）新词情感倾向的确定。 本文通过 word2vec 模型来确定这些新词的情感倾向。word2vec(word to vector)是 Google 于 2013 年开发的一个向量表示工具，通过训练，可以用 K 维向量空间来表示微博文本，然后对向量进行运算，得到空间上的相似度，也就计算出微博文本语义上的相似度。

因此利用 word2vec 模型的这个功能，可以计算出微博文本中与新词最相似的那个词语，然后再去匹配基础情感词典，确定相似词是正面的，负面的，还是中立的，从而得到新词的情感倾向。最终将正负面情感倾向的新词加入到情感词典中。

1.4　特征选择

采用条件随机场识别的情感要素包括：正负面情感词和复杂结构词，正负面情感词是决定微博句子情感倾向的主要因素，复杂结构词是影响微博句子真实情感倾向的主要因素，二者缺一不可。准确识别这些情感要素离不开有价值可靠的语言特征。本文选取四类特征来识别情感要素：基本特征，语义依存特征，复杂句式特征，情感特征，其中基本特征和语义依存特征都是通过哈尔滨工业大学的语言技术平台（language technology platform,LTP）获得的，复杂句式特征通过构建的复杂结构词表获得，情感特征通过扩展的情感词典获得。具体如下：

基本特征包括词和词性特征。词特征，即分词后微博文本中的每个词，是细粒度情感分析的具体对象。词性特征，指分词后每个词的词性，如正负面情感词一般是形容词，复杂结构词一般是连词，所以提取每个词的词性特征对情感要素的提取有一定的辅助作用[13]。

复杂句式特征，指影响微博文本情感倾向的一些复杂结构词，主要包括转折结构词和否定结构词。转折结构词即微博文本中的转折连词，比如：“周末好开心啊，但突然想到一堆作业没写，伤心中……”，此文本中的一个转折词“但”将文本整体情感完全逆转，如果不考虑此类结构词，很难判断文本的真正情感倾向；故根据学生微博特点构建如表 2 所示的转折词表。

表 2　转折词表

类别	举例
转折词	虽然……但是、然而、但是、却、 不管（无论）……也（都）、 尽管……但、即便……也、即使……也

否定结构词即微博文本中的否定词，修饰情感词，并影响微博的情感倾向，如：“早就预料到了，所以一点都不难过”。“难过”是负面情感词，在否定词“不”的修饰下，微博文本情感倾向呈正面的，忽略此类结构的文本将丢失大量有价值的文本信息。故构建如表 3 所示的否定词表。

表 3　否定词表

类别	举例
否定词	不、不是、没有、不要、别、无、不太

将上述构建的转折结构和否定结构词表，分别与分词后的词语进行匹配来判断该词是否具有复杂句式特征。该特征采用二元特征值表示法如表 4 所示。

表 4　复杂句式特征

特征	特征信息	表示方法
复杂句式特征	转折词或否定词	用 1 表示
	其他词	用 0 表示

情感特征，为了识别出的正/负面情感词更加准确，通过上述构建的情感词典与分词后的各词进行匹配，将匹配正确的情感词，采用三元特征值表示法表示，具体如表 5 所示。

表 5　情感特征

特征	特征信息	表示方法
情感特征	正面情感词	用 1 表示
	负面情感词	用-1 表示
	其他词	用 0 表示

语义依存特征，可以将各个词语之间的语义关联以一种依存关系的结构呈现出来，不受句法结构的影响，这是跟依存句法特征[13]最大的区别。在语义依存特征的辅助下，大大提升正负面情感词和复杂结构词识别的准确率。

如句子 1 “不是这样的，我很开心。” 这条微博中，“不”这个词属于否定结构词，“开心”是正面情感词，如果仅靠复杂句式特征，那“不”就会被标志为复杂结构词，从而影响微博文本

的真实情感倾向。融合语义依存特征后，根据各个词之间的语义关联，可分析出“不”并不是修饰“开心”的复杂结构词，如图 1 所示，经过语义依存分析后，可看出“不”和“开心”之间并没有语义关联，而“很”与“开心”之间是“mDegr”程度标记的关系。

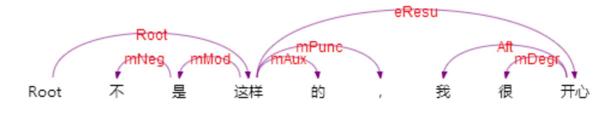


图 1　句子 1 的语义依存分析图

从而在最终的识别结果中，“不”会被识别为背景词，这样，根据识别出的正面情感词，即可判断出微博文本的真实情感倾向。同理，正负面情感词的识别亦是如此，如句子 2：好吧，这么差的饭店以后不会来了。依赖情感特征捕捉情感词，“好”和“差”都会被识别为情感词，融合语义依存特征后，即可判断出“好”是语气词，“差”才是真正带有感情色彩的形容“饭店”的情感词。如图 2 所示，经过语义依存分析后，将分词后的词之间的关系都展现出来，可以看到“好”与“吧”之间的关系是“mTone”，表示语气标记，而“差”才是真正的对“饭店”的评价，呈现出“Desc”描写的关系。

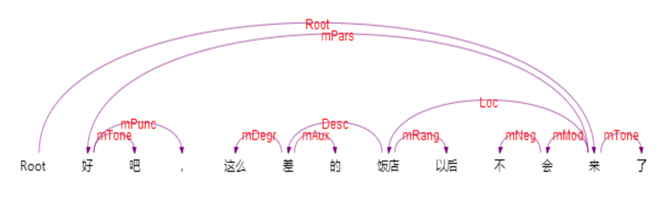


图 2　句子 2 的语义依存分析

这样引入语义依存特征，对准确识别情感要素有很大帮助，而且比依存句法特征价值更高，取得效果也更好。此特征的详细描述如表 6 所示。

表 6　语义依存特征

特征	特征信息	含义
语义依存特征	父节点词	分词后的各词在依存关系中的父节点词
	父节点词的词性	分词后各词的父节点词的词性
	依存关系	分词后的各词与其父节点词之间的依存关系

2　基于复杂句式的微博情感倾向分析

情感要素抽取之后，也就实现了细粒度的情感分析，借助抽取出的情感要素，还需要分析微博文本的情感倾向。一般的微博文本情感分析属于粗粒度的情感分析，经常用到的方法是情感词典法和机器学习法，这类方法缺乏对复杂句式的有效分析，即没有考虑到微博文本中的复杂结构词。因此提出基于复杂句式的情感倾向分析方法。

2.1　复杂句式文本介绍

根据不同的句子结构，中文文本呈现不同的句式。一般来说中文文本有简单结构和复杂结构两类。微博文本亦是如此，简单结构的文本是指由主谓或主谓宾结构构成的句子，这些句子结构简单，表达单一，统称为简单句式，而复杂结构的文本是指由两个或两个以上单句组成，单句之间用分号、逗号等标点符号隔开；或通过特定的连接词连接的句子，这类句子之间相互关联，合在一起才能表达完整意思，称为复杂句式[14]。如：“今天天气很好”这是一个简单句式，而“今天天气很好，但心情却不美好”这就是一个复杂句式，其中包含连接词“但”，侧重强调“心情不好”。

根据文本中单句之间的关系，复杂句式有八种结构：选择结构、并列结构、递进结构、条件结构、转折结构、取舍结构、因果结构、假设结构[15]。具体如表 7 所示。

表 7　复杂结构举例

复杂结构	常用连接词举例
并列结构	又……又、那么……那么
递进结构	不但……而且、不仅……还
转折结构	虽然……但是、……但、即使……也
条件结构	只要……就、只有……才
因果结构	因为……所以、因此
选择结构	不是……就是、是……还是、或者……或者
取舍结构	与其……不如、宁可……也
假设结构	如果……就

现在的微博文本表达形式多样，语言随意灵活，多为复杂句式，因此分析微博文本的情感倾向，更应该考虑微博文本中的复杂结构。通过分析以上 8 种复杂结构，可知微博文本中的转折结构词，会影响文本的情感倾向，如“天气虽好，但心情却很低落”这条微博文本中，有两个情感词，一个“好”，一个“低落”，转折结构词“虽……但”使文本情感发生转变，这种情感转变叫做情感偏移。除了因单句之间的关系形成的复杂结构，本文将否定结构的微博文本也归为复杂结构，因为否定结构也会引起微博文本情感偏移，如：“我不太高兴”，“高兴”是正面的情感词，否定词“不”的修饰使句子表现出负面情感，因此在判断微博文本的情感倾向时，必须考虑到文本中的转折结构词和否定结构词。

2.2　情感倾向性分析

基于复杂句式的文本情感倾向性分析，需要将识别出的情感要素与文本模式相结合。文本模式有如下三种：

a)无复杂结构词模式，此类模式的文本情感倾向与情感词极性一致。

b)否定结构词+情感词模式，此类模式的文本情感倾向与情感词极性相反。

c)转折结构词+情感词模式。转折结构词又分单个关联词模式和多个关联词模式。单个关联词模式，如“天气很好，但却有点难过”，此时文本的情感倾向性跟转折结构词后的情感词极性一致；多个关联词模式，如“虽然天气很好，但有点难过”，此

时文本的情感倾向性与第二个关联词后的情感词极性一致。

具体文本情感倾向的分析过程如图 3 所示。

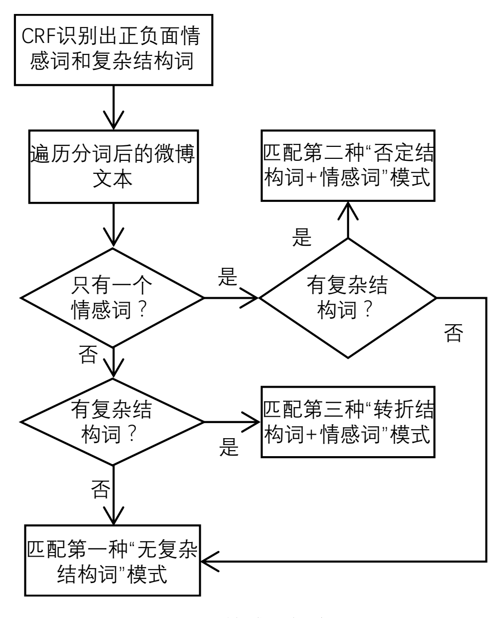


图 3　微博情感分析流程图

通过上述方式可分析出每条微博的情感倾向。当匹配第一种模式时，文本情感倾向与情感词极性相同。当匹配第二种模式时，文本情感倾向与情感词极性相反。当匹配第三种模式时，转折关联词只有一个时，文本情感倾向与关联词后的情感词极性相同；转折关联词有多个时，文本情感倾向与第二个关联词后的情感词极性相同。最后，统计判断出的正负面情感的微博文本数，得到学生整体的情感倾向。

3　实验结果及分析

3.1　语料收集及预处理

本次实验的数据来自新浪微博，以"山西农业大学"为关键字爬取学生微博 10000 条，经过人工分类，具体的数据信息如表 8 所示。

表 8　微博文本统计信息

微博	微博数
正面情感	5 002
负面情感	2 891
中立	2 107

首先去除中立的微博文本，对剩余的含有无效网址及字符的微博文本进行预处理，然后，对经过分词后的微博文本采用文献[13]提出的协同训练的方式进行半自主标注，最后进行五折交叉验证实验，即将所有的微博文本分为五份，其中四份为训练集，用于训练分类模型，一份为测试集，用于验证情感要素识别的效果。

3.2　情感词典构建结果

根据 1.3 节扩展情感词典的方法，在基础情感词典 HowNet 的基础上，加入了当下较流行的网络词汇，如下表 9 所示。

表 9　常用网络词汇

类别	举例	数量
正面情感词	笔芯，比心，么么哒，棒棒哒等	20
负面情感词	蓝瘦，香菇，扎心，智障等	20

3.3　情感要素识别的结果

实验中采用 CRF 模型作为情感要素识别的工具，采用哈工大的 LTP 语言云获得词、词性及语义依存特征，采用构建的转折结构词表和否定结构词表获得复杂句式特征，采用扩展的 Hownet 的中文情感词典获得情感信息特征。选用不同的方法对微博语料进行情感要素识别效果的比较，将精准率 P(Precision)，召回率 R(Recall)和 F-measure（精准率和召回率的调和平均值）作为识别效果的评价指标。得到结果如表 10 所示。

表 10　情感要素识别

方法	特征选取	情感要素	学生微博文本 P(%)	R(%)	F(%)
1	词+词性+复杂句式特征	复杂结构词	78.6	72.0	75.2
		正面情感词	71.2	59.6	64.9
		负面情感词	69.5	55.8	61.9
2	词+词性+情感特征	复杂结构词	62.3	60.7	61.5
		正面情感词	78.4	76.2	77.3
		负面情感词	80.7	68.9	74.3
3	词+词性+复杂句式特征+情感特征	复杂结构词	82.8	74.5	78.4
		正面情感词	79.5	80.1	79.8
		负面情感词	76.3	74.9	75.6
4	词+词性+复杂句式特征+情感特征+依存句法特征	复杂结构词	85.3	80.5	82.8
		正面情感词	83.6	79.4	81.4
		负面情感词	85.3	83.2	84.2
5	词+词性+复杂句式特征+情感特征+语义依存特征	复杂结构词	85.1	87.8	86.4
		正面情感词	90.6	86.9	88.7
		负面情感词	87.7	92.6	90.1

从表中可以看出，第一种方法选取了词、词性和复杂句式特征，在这三种特征的作用下，复杂结构词的精准率和召回率都达到了 70%以上，但是情感词的综合识别率却不是很高。主要因为没有可以捕捉情感词的具体特征；第二种方法选取了词、词性和情感特征，情感词的综合识别率提高了，但复杂结构词的精准率和召回率却大幅度下降，只有 60%多。因此有了第三种方法，结合了前面两种方法的特征，复杂句式特征对捕捉复杂结构词有一定的作用，情感特征对捕捉情感词有一定的作用，故大大提升了情感要素的识别率。第四种方法加入了依存句法特征，可分析文本中各词之间的依存关系并揭示句法结构，可发现，情感要素的识别率都明显提升了，主要是因为依存句法特征有效揭示了各词之间的依赖关系，排除了一些没有情感意

义的情感要素。第五种方法，将依存句法特征换成语义依存特征，不同点是语义依存特征能跨越句子表层句法结构的约束，获取深层的语义信息，得到的效果比依存句法特征好很多，精准率和召回率都有很大提升。

3.4 情感倾向分析结果

识别情感要素后，基于复杂句式对微博进行情感倾向性分析，并与传统的朴素贝叶斯分类方法进行比较，结果如表 11 所示。

表 11 微博情感倾向分析结果

序号	方法	正面情感			负面情感		
		P/%	R/%	F/%	P/%	R/%	F/%
1	朴素贝叶斯	79.2	72.1	75.5	73.4	69.7	71.5
2	CRF+复杂句式	88.3	85.6	86.9	84.9	90.8	87.8

分析结果,可发现朴素贝叶斯分类器分类得到的情感倾向结果,综合识别率要偏低,主要是因为没有选取比较有效的语言特征,对文本中的复杂句式也没有处理。针对这些问题,本文提出结合 CRF 和复杂句式的跨粒度情感分析方法,通过 CRF 模型识别出真正影响微博情感的情感要素,再结合复杂句式判断每条微博的情感倾向,这样判断出的正/负面情感文本,精准率和召回率都达到 85%左右。

4 结束语

现有的情感分析方法大都是粗粒度的分析方法，分析一条微博文本的情感倾向，忽略了微博文本中的复杂句式，没有考虑影响情感倾向的各种语言特征，本文提出一种跨粒度的情感分析方法，首先利用条件随机场模型，充分考虑文本中的各种语言特征，对微博文本进行细粒度的情感分析，识别出文本中的情感要素，然后结合复杂句式，判断文本的情感倾向，实现粗粒度的情感分析，最终得到学生微博整体的情感倾向，达到了从细粒度到粗粒度的跨粒度情感分析。

本文实现了对学生微博的情感分析，即通过分析得到了学生整体的情感倾向，便于学校管理和及时了解学生情绪动态。但是没有做细致分析来了解学生情感的对象，未来的研究工作将侧重于分析学生情感的对象，引出学生比较关注的热点话题以及对此话题的观点，达到舆情分析的目的，最终还可以进行相关的舆论引导，有效遏制不良舆情的发展。

参考文献：

[1] 肖江，丁星，何荣杰．基于领域情感词典的中文微博情感分析 [J]. 电子设计工程, 2015, 2015 (12): 18-21.

[2] 陈晓东．基于情感词典的中文微博情感倾向分析研究 [D]. 武汉：华中科技大学, 2012.

[3] Catal C, Nangir M. A sentiment classification model based on multiple classifiers [J]. Applied Soft Computing, 2017, 50 (2017): 135-141.

[4] Liu S, Li F, Li F, *et al.* Adaptive co-training SVM for sentiment classification on tweets [C]// Proc of ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2013: 2079-2088.

[5] Hu M. , Liu B. Mining and summarizing customer reviews [C]// Proc of KDD. New York: ACM Press, 2004: 168-177.

[6] 赵文婧．产品描述词及情感词抽取模式的研究 [D]. 北京：北京邮电大学, 2010.

[7] Xu R, Wong K F, Xia Y. Opinmine–opinion analysis system by CUHK for NTCIR-6 pilot task [C]// Proc of NTCIR. 2007: 350-357.

[8] 李阳辉，谢明，易阳．基于深度学习的社交网络平台细粒度情感分析 [J]. 计算机应用研究, 2017, 34 (3): 743-747.

[9] Jin W, Ho H H. A novel lexicalized HMM-based learning framework for web opinion mining [C]// Proc of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 465-472.

[10] Liu P, Meng H. SeemGo: Conditional random fields labeling and maximum entropy classification for aspect based sentiment analysis [C]// Proc of International Workshop on Semantic Evaluation. 2014: 527-531.

[11] Lafferty, A. McCallum, F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishing, 2001: 282-289

[12] 赵龙，郭立，谢锦生．条件随机场模型的场景描述 [J]. 中国图象图形学报, 2013, 18 (3): 26-31.

[13] 刘丽，王永恒，韦航．面向产品评论的细粒度情感分析 [J]. 计算机应用, 2015, 35 (12): 3481-3486.

[14] 郄鹏，李爱萍，段利国．基于转折句式的文本情感倾向性分析 [J]. 计算机工程与设计, 2014, 2014 (12): 4289-4295.

[15] 郄鹏，段利国．基于复杂句式的文本情感倾向性分析 [J]. 计算机应用与软件, 2015, 32 (11): 57-61.